MLNS – Final Project Report

Analyzing global air traffic

Victor MAILLOT - B00803877
Marina PELLET - B00794582
Emma RIGUIDEL - B00798799
Emma SERFATY - B00794569

Abstract

In this paper, we present our exploratory and predictive analysis on a worldwide airports network. Our network is a weighted graph, with nodes representing airports, edges representing connections between airports, and edge weights representing the number of flights between them. Our analysis revealed that only a few airports dominate the network, capturing a large proportion of the flights. Indeed, using state-of-the-art centrality measures, we found that European and US airports dominate the top 10 ranking of most important airports, with European airports acting as bridges, ranking high in terms of betweenness centrality, while US airports are better connected to each other, proven by a higher closeness and degree centrality. Additionally, we found that Asian airports rank lower in terms of eigenvector centrality, indicating that they are poorly connected to the most influential airports. Airports in Oceania, South America, and Africa are relatively isolated from the main international airports, but some still play important roles as bridges, shown by the higher betweenness centrality rankings. The clustering work presented similar results, showing a strong dominance of a few globally important airports alongside other relatively isolated groups. Finally, we used the existing network to build a link prediction model that showed very good results, and indentified the pairs of unconnected airports that are the most likely to be connected in the future.

1 Introduction

A city's air traffic is a crucial indicator of its size and characteristics. A thorough examination of air routes is necessary to enable effective resource allocation and optimize the advantages of air traffic. To effectively allocate resources and determine the quickest routes for passengers, we assess the many international air routes in this project and identify the most crucial routes to open. Our approach entails representing the issue as a graph, where nodes stand in for airports and edges for the connections between them. Using the NetworkX package, we will convert our dataset into a graph by weighing the edges according to the quantity of flights that pass between the airports. Then, through an exhaustive explanatory analysis, we will pinpoint significant graph properties to understand the network better. To distinguish between different groups of airports, we intend to employ community detection. To determine the most practical new routes, we will investigate link prediction.

2 Problem definition

The first goal of this project is to perform an exploratory analysis, especially by applying centrality measures to point out the most important airports and connections in our air traffic network. To do so, we'll use density and several centrality measures:

- Density of a graph:

$$D = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V|-1)}$$

With D the density of a undirected graph, |E| the number of edges, |V| the number of nodes.

- Degree centrality:

$C_d(i) = k(i)$, where $k(i)$ is the degree of node i

The normalization is done by dividing the results by number of nodes.

- Closeness centrality:

$$C_{cl}(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$

$d(i, j)$ is the length of the shortest path between i and j

It measures the ability to quickly access or pass information through the graph. The closer to one the value, the more influential the node.

- Betweenness centrality:

$$C_{bt}(i) = \sum_{s \neq i \neq t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}$$

– $\sigma(s, t)$ is the total number of shortest paths from s to t
– $\sigma(s, t \mid i)$ is the number of shortest paths from s to t that pass through i

According to this approach, a node is important if it lies in many shortest paths. This centrality highlights essential nodes in passing information through the network.

- Eigenvector centrality:

$$x_i = \frac{1}{\lambda_1} \sum_j \mathbf{A}_{ij} x_j$$

This measure states that the importance of a node increases if it is connected to other important nodes. The eigenvector centrality of a node can be large either if it has many neighbors-, or if it has important neighbors.

3 Related Work

The paper "*New centrality and causality metrics assessing air traffic network interactions*" by P. Mazzarisi et al. (2018) was previous research relevant to our project. This paper discusses the application of centrality measures to analyze the structure of air traffic networks. The authors use degree, betweenness, and eigenvector centrality measures to identify the most important airports in the network, and they show that these measures are correlated with airport traffic volume.

"*A study of the U.S. domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016*" by Leonidas Siozos-Rousoulis, Dimitri Robert, Wouter Verbeke (2019) is a research paper that examines the U.S. domestic air transportation network from 2001 to 2016. The authors analyzed the network's topology and robustness by studying its structural properties, such as degree distribution, average path length, and clustering coefficient. They found that the network has become more connected and centralized over time, with larger hubs playing a more important role. The network has become more robust to random failures but less robust to targeted attacks. The findings of this study provide insights into the evolution of air transportation networks over time and could be useful for understanding the resilience of airport networks to disruptions.

Another paper titled "*Link prediction and clustering in airport networks: a comparative study*" by A. Solé-Ribalta et al. (2018) investigates the use of link prediction algorithms and clustering methods in airport networks. The authors compare different link prediction algorithms, including the Adamic-Adar and preferential attachment algorithms, and they use clustering methods such as the Louvain algorithm to identify groups of airports with similar properties. This paper could be relevant to our project as it provides a comparison of different link prediction and clustering methods in the context of airport networks. These papers could help us understand deeper what is at stake in our network analysis, giving us some context and some insights regarding air network traffic. We also wish to test other centrality measures or link predictions algorithms approaches on our network.

4 Methodology & Evaluation

Data preprocessing

The initial step of this project involved converting our dataset, which was in csv format, into a graph format. Indeed the focus of our project revolves around graph representations of the data. Therefore, the first task is to convert the table into a graph where the nodes represent airports and the edges indicate the existence of a route between two airports.

Our initial dataset consisted of 9 columns: `airline`, `airline_id`, `source_airport`, `source_airport id`, `destination_airport`, `destination_airport id`, `code_share`, `stops`, `equipment`. It had 67,663 rows. We started by removing null values for

`destination_airport_id` and `source_airport_id`.We noticed that the dataset contained only 11 lines with the value of `stops` higher than 1, and we removed these lines to be thorough in our analysis.

After that, we performed some aggregations with pandas to obtain our final dataset and create the network. To do so, we only needed three columns in the dataset: the source airport, the destination airport, and the number of flights between them. After the aggregation, the number of rows in the dataset was 37,274 rows, corresponding to all possible pairs of airports with flights between both. However, we quickly noticed that the number of flights going from airport A to airport B was the same as the number of flights going from airport B to airport A, almost every time. That is why we decided to keep only the sum of the number of flights between two airports, converting the network from a directed to an undirected graph. At the end, our dataset from which we created the graph contained three columns: `airport A`, `airport B`, `total_flights`. This final dataset contained 19,080 rows.

After converting the dataframe into a graph, we end up with a network in which a node corresponds to an airport, and the edges are weighted based on the number of flights between two airports. The graph is undirected, as the value of edge corresponds to the sum of flights between airports A and B.

Characteristics of our network

The start off our exploratory analysis we computed some classic indicators:

Feature	Measure	Value
Nodes	Count	3330
Edges	Count	19,080
Degree	Min	1
	Max	248
	Mean	11.46
	Median	3
Weight	Min	1
	Max	39
	Mean	3.52
	Median	2
Weighted degree	Min	1
	Max	1,826
	Mean	40.37
	Median	8
Connected components		7
Density		0.0034

Figure 1: Main characteristics of the network

This information gives us important information to start our analysis. The observed density of 0.034 is quite low, indicating that only a few airports are linked to each other. The metrics concerning degrees and weights of the graph enable us to understand more about the network configuration and the reason why density is low. Indeed, on average, an airport is linked to 11.46 other airports, and the median number of connections of an airport is 3. The most connected airport only has 248 connections, which is relatively low, knowing that there are 3,330 airports in total.

Figure 2 clearly shows that the degree distribution is long tailed, which means that most airports are connected to only a few airports: 78% to less than 9, against 1.9% to more than 100. Only a tiny percentage of them are big airports that influence the average number of connections. In *Figure 2*, we only showed data with degree less than 60 so that it could be readable.

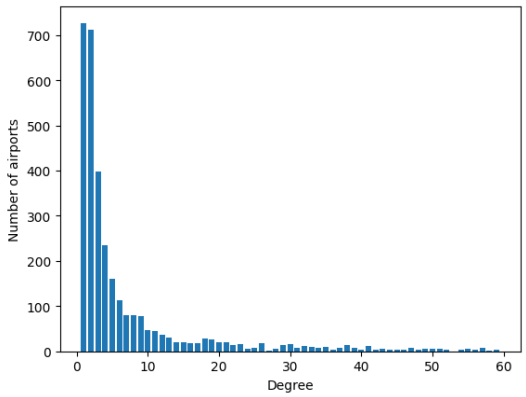

Figure 2: Histogram of airports relatively to their degree (degree < 60)

As our graph takes into account not only the existence of connections between airports, but also the number of flights between them (weight of the edges), we then decided to look into the distribution taking into account the weight of the edges. In *Figure 1*, we can see that the maximum number of flights between two airports is 39, the average 3.52, and the median 2, which is quite low as the dataset focuses on only 1 small period of time t.

We also looked into the distribution of the weighted degree, meaning the number of flights passing through the airports. On average, there are 40 flights passing through an airport, but the median being 8, in practise most airports have a limited activity. The distribution of weighted degree is even more extreme than the distribution of the degree, indeed for 78% of airports, there are less than 30 flights passing through them in the dataset, whereas 1.35% of airports host more than 500 flights.

To complete our overview of the network, we looked into the number of components of the graph. We found out that the graph contains 7 components, including one with 3,304 nodes, meaning 99.2% of total airports. Thus, other nodes are very isolated. (Figure 3)

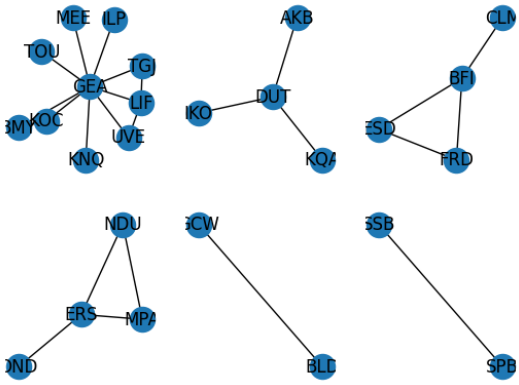

Figure 3: The 6 isolated components of our graph

Each of the 6 six components corresponds to an isolated region of the world: Nouméa island on the top left, Namibia, Alaska, Grand Canyon area, and Virgin Islands, where airports are very small.

After this first exploratory phase, we explored node centrality using 4 metrics: degree, closeness, betweenness, and eigenvector centrality, each of them giving us different insights about the airports present in the network.

Centrality measures

- *Degree centrality:*

To detect the most important airports, we started by using unweighted degree, the simplest centrality measure. Airports with high degree centrality are directly connected to many other airports in the network, and thus play an important role in connecting different parts of the network.

Here is the top 10 of the airports which have the highest number of connections and thus the highest degree centrality:

Airport	City	Degree
AMS	Amsterdam	248
FRA	Frankfurt	244
CDG	Paris	240
IST	Istanbul	234
ATL	Atlanta	217

Figure 4: Top 5 airports based on their unweighted degree centrality ranking

Unsurprisingly, they are some of the world's biggest cities' airports. The fact that these airports are connected to the highest number of airports suggests that they are major hubs in the global air transportation network. Therefore, they play a critical role in connecting different regions of the world. After normalization, the top 10 airports have a degree centrality between 16.3 and 21.6, higher than the average number of connections by airport. There are also 11 airports being connected to 1 airport only; their degree centrality is 0.08726, a value 11.46 times smaller than the average degree by airport, 11.46.

If we now consider the weights of the edges (meaning the total number of flights from and to that airport) to display the most important airports in terms of connections, the ranking somewhat changes:

Airport	City	Total flights
ATL	Atlanta	1826
ORD	Chicago	1108
PEK	Beijing	1069
LHR	London	1047
CDG	Paris	1041

Figure 5: Top 5 airports based on their weighted degree centrality ranking

Taking into account the weight of edges is particularly useful here in identifying major hubs. Here, we notice that the top 2 is composed of 2 American cities, Atlanta and Orlando, respectively hubs of united airlines, american airlines, and delta airlines, the three biggest airlines companies in the world. London, Los Angeles and New York also appear in the top 10, which means that their airports have a high number of flights, even if they are not well-connected in terms of the number of directly connected airports. We can guess that various factors such as location, airline partnerships, or route demand can contribute to this.

For other centralities measures, we have decided to use weights of edges for computation, as it gives a better representation of the reality. For each of them, we have used a function to obtain the centrality for each node, and then we called a function that we created, *sort_centrality* to display results.

- *Closeness centrality*

In our context, closeness centrality measures the average length of the shortest path between an airport and all other airports in the network. Airports with high closeness centrality are closer to other airports in the network and can be considered more accessible and influential.

Using the closeness centrality measure, we noticed that the Top 5 most important airports are in Frankfurt, Paris, London, Amsterdam and Dubaï, which means that they are more central, i.e., that they are more easily reachable from all other airports. Atlanta, Chicago and Beijing are the top 3 cities in terms of degree centrality, which focus on the number of direct connections, we can see here that they are not in the top 5 of closeness centrality that looks at the overall position of a node in the network and how easily it can access other nodes.

- *Betweenness centrality*

Betweenness centrality measures how often an airport lies on the shortest path between two other airports in the network. Airports with high betweenness centrality serve as bridges or connectors between different parts of the network and are critical for maintaining the network's overall connectivity. After performing our analysis, we noticed that the most important airport using this centrality measure is in Alaska, in Anchorage city. This airport is indeed the hub for Alaska Airlines. It is also the 4th largest airport in the world in terms of commercial traffic. The 4 other most important airports are Los Angeles, Paris, Dubai and Frankfurt. Contrary to closeness centrality which measures how fast a node can reach other nodes, betweenness centrality measures how important a node is for connecting different parts of the network. Even if we notice that some airports are on top of both rankings, betweenness and centrality enables us to discover some interesting points. Indeed, we can see that Los Angeles and Anchorage are crucial to maintain an efficient airport network. Additionally, in the top 15, we also noted the presence of airports with a much lower ranking using other measures of centrality, located in cities such as Sao Paulo, Sydney, Brisbane, Bogota and Johannesburg. These airports, located on continents far from the core network, may be valuable ‘bridges’ that connect the continents together.

- *Eigenvector centrality*

Eigenvector centrality measures an airport's influence based on the influence of its neighboring airports. Airports with high eigenvector centrality are connected to other highly influential airports in the network and can be considered more important in terms of their overall influence within the network.

Using this centrality measure, most important airports are located in Amsterdam, Frankfurt, Paris, Munich, Roma and London. We also noticed highly ranked airports that are not usually so high in the ranking, all of them in European cities: Barcelona, Zurich, Madrid, Brussels, Dublin, Manchester, Copenhague, Vienna… All of them may be

well connected to influential nodes in the network, i.e. hubs located in cities such as Frankfurt, Amsterdam, Paris.

- *Sum up centralities measures*

Airport	City	Median ranking	Hub of
CDG	Paris	3	Air France
LHR	London	4	British Airways
FRA	Frankfurt	5	Lufthansa
DXB	Dubaï	5	Emirates Airlines
LAX	Los Angeles	6	Alaska Airlines, American airlines
AMS	Amsterdam	6	KLM
PEK	Pékin	7	Air China
ORD	Chicago	8	United Airlines
JFK	New York	9	American airlines
YYZ	Toronto	10	Air Canada

Figure 6: Top 10 airports based on the median of their ranking for 4 centrality measures

To compute the final ranking, we used the ranks of all previous centrality measures. To give an example, CDG airport has the 5th highest degree centrality, the 2nd highest closeness centrality, the 3th highest betweenness centrality and the 3rd highest eigenvector centrality. Finally the median of these rankings is 3, which makes Paris CDG airport the most important airport of the network.

According to this ranking, the 5 most important airports of this dataset are located in Paris, London, Frankfurt, Dubaï and Los Angeles. Each of them are the main hub of an important airline company. More specifically, apart from CDG airport whose rank is always in the top 5 for all centrality measure, we notice that LHR betweenness rank is 13, which means that this airport is not as important as others for flight connection. DXB airport is slightly less important in terms of degree and eigenvector centralities, while LAX eigenvector ranking is only 49, which means that this airport is not very well connected to influential nodes.

Overall, these 10 airports (Figure 6) are almost everytime the main hub for the world's most important company. At first glance, it may seem surprising to see airports ranked so high in cities such as Amsterdam, Munich, Chicago, Atlanta or Dallas, but these airports are in fact very strategic because they are hubs for very important international airlines, especially for the last two American cities, the main hubs of Delta Airlines and American Airlines, the two largest airlines in the world.

This ranking also shows the predominance of European and US airports in international air traffic, especially in the top 10. In particular, it shows that European airports are generally at the top of the ranking in terms of betweenness suggesting they act as important bridges, while US airports are better connected to each other, which translates into greater closeness and degree centrality. Asian airports, on the other hand, are a little further down the rankings, partly because of eigenvector centrality, as they are poorly connected to the most influential airports. Finally, the airports of Oceania, South America and Africa are quite isolated from the main international air networks. It is not until the 48th position in the ranking that an airport from one of these continents appears, located in Sao Paulo. However, some of them are still important as they act as bridges, resulting in higher betweenness centrality ranking.

Clustering

Let’s now identify the different groups of airports within the world traffic. To do this, we will perform spectral clustering which is a popular algorithm in the task of clustering on a graph. The idea of this method is to use the eigenvalues of the similarity matrix and perform a dimension reduction before finally determining the different clusters of the graph. Despite its overall effectiveness, spectral clustering has a disadvantage: it is necessary to know in advance the number of clusters that we want to determine. Knowing that we are studying global air traffic network, we had the idea of setting the number of clusters to 5 in order to see if the algorithm would sort airports by continent. This could make sense because one can assume that airports located in the same continent have more trips in common.

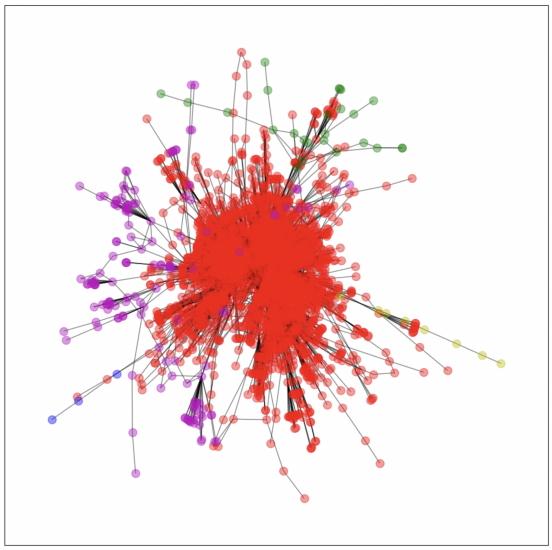

Figure 7: Output of Spectral Clustering (k=5) on the main connected component of the data

From *Figure 7*, we can observe that there is one main cluster (in red) containing 3117 airports, and four other groups containing each between 151 and 3 airports. First, we can conclude that our idea falls through and that the clusters do not really resemble the 5 continents. In order to understand the results of the clustering and to understand which are the most important airports in each group, we calculated the closeness centrality of each airport within its cluster. (*Figure 8*)

Cluster	Airport 1	Airport 2	Airport 3
1 - Europe (red, 3317 airports)	FRA (0.422)	CDG (0.418)	LHR (0.416)
2 - Brazil (blue, 3 airports)	SXO (1.0)	STZ (0.666)	GRP (0.666)
3 - Greenland (green, 25 airports)	GOH (0.511)	SFJ (0.471)	JAV (0.421)
4 - North America (magenta, 151 airports)	ANC (0.498)	BET (0.393)	FAI (0.382)
5 - US - Australia (yellow, 8 airports)	FBS (0.438)	RCE (0.438)	LKE (0.389)

Figure 8: Top 3 airports (by closeness centrality) by cluster

In view of the different sizes of heterogeneous clusters, it can be said that world air traffic is dominated by a highly connected group of airports, the three largest of which are European (cluster 1). Cluster 2 contains three Brazilian airports which therefore seem to be isolated from world traffic. Cluster 3 is dominated by three airports in Greenland and mainly includes airports in this country. We can also identify a cluster centered on the United States and Alaska comprising 151 airports in North America (cluster 4). The last cluster (cluster 5) also includes airports in the United States as well as some present in Australia, there is also the strong presence of airports located in the state of Washington.

Link prediction

An airport is connected with another if there has been at least one trip between these two. It would now be interesting to see if we can train a model capable of predicting potential future trips between airports. To do so, we have chosen the Supervised Link Prediction method which makes it possible to determine the most likely future connections between airports. This method breaks down into three steps: first, randomly separate the dataset into a training set and a test set, then train a model on a binary classification task (the link exists or the link does not exist between two airports) and finally to evaluate its performance on the test data set using a ROC curve comparing the True Positive rate with the False Positive rate. After keeping 60% of the dataset for training, choosing basic features (degree centrality, betweeness centrality, preferential attachment, Adamic Adar index, Jaccard coefficient) to characterize each airport, and fitting a Logistic Regression model to the data, we obtained an AUC (Area Under the Curve) of 0.973, showing that our model is able to predict the occurrence of a new link very accurately.

After training our model, we ran it on our network's non-edges, i.e. non-connected airports. The model predicted connections for around 6% of airport pairs, for the 5.4 million non-connected pairs. Based on our predictions, the following new connections are the most likely to open:

- ARN (Stockholm) - MRS (Marseille)
- NRT (Tokyo Narita) - MCO (Orlando)
- PHL (Philadelphia) - NCE (Nice)

- NRT (Tokyo Narita) - XIY (Xi'an)
- WAW (Warsaw) - SAW (Istanbul)
- MCO (Orlando) - BCN (Barcelona)

It is worth to note that that some of these connections do exist (e.g.: Warsaw - Istanbul), while others do not (Stockholm - Marseille). It might be due to the fact that our dataset is from a few years ago, or that it may be incomplete.

6 Conclusions

The goal of our project was to identify the most important airports of the network using different criteria, and then to understand its structure better by doing some clustering and link prediction. We started by conducting an exploratory analysis which revealed that only a few airports are dominating the network, capturing a huge part of the flights. To define the importance of a node, we then used four centrality measures to rank the importance of airports in a network. After analyzing the data, we concluded that European and US airports dominate the top 10 rankings, with Paris, London, Frankfurt, Dubai, and Los Angeles being the top five ranked airports. These airports are main hubs for important airline companies. Additionally, we found that Asian airports rank lower in terms of eigenvector centrality, while airports in Oceania, South America, and Africa are relatively isolated from the main international air networks, but some still play important roles as bridges, resulting in higher betweenness centrality rankings. From the clustering analysis, we found that there is a central group of airports, dominated by European strategic hubs, while four other groups stood out, located in Brazil, Greenland, North America and Australia. These groups confirms that the network of airports is heterogeneous, and that flights seem to concentrate in regions of the world. Finally, from the link prediction task, we were able to identify 6% of non-existing connections which are the most likely to open, or that should be a priority when opening new routes. This model could be useful to effectively allocate resources and serve the customers best when expanding the network. In future studies, it could be interesting to analyse the impact of an airport's failure in the world (due to for example political instability, power shortage, etc) and to include the cost component. This type of analysis could improve the resilience of the network by ensuring alternative routes are built and appropriate measures are taken to prevent such failures.

References

- *New centrality and causality metrics assessing air traffic network interactions*" by P. Mazzarisi et al. (2018)
- "*A study of the U.S. domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016*" by Leonidas Siozos-Rousoulis, Dimitri Robert, Wouter Verbeke (2019)
- "*Link prediction and clustering in airport networks: a comparative study*" by A. Solé-Ribalta et al. (2018)
- *Air traffic network analysis* by Vinicius, M. Available at: https://medium.com/@mvinnicius22/air-traffic-network-analysis-976ad48f048c